Grammatical variation in English is not so difficult

Thomas Van Hoey
Matt Hunt Gardner
Benedikt Szmrecsanyi

ISLE 7
22/06/2023
University of Queensland

Variable grammars

I don't know I think v05 v08 they'll go I think v05 v08 they'll they're good but they’re v13 not great but I think v05 v08 they’ll

(female / Western / born 1956 in Switchboard corpus)

V05	choice between	that / __zero__ complementizer
V08	choice between	will / going to / shall / …
V13	choice between	V not / Vn’t

Why variation matters

> “Among the achievements of this past century is the discovery of the incredible variability in human language [...] This variability continues to present a set of fundamental puzzles that need to be solved to find the key in explaining and understanding variability as an inherent property of human language.
>
> (van Hout & Muysken 2016:250-251)

Alternation variation at ISLE 7

❖ dative	*ditrans* | *prep*	Paolini
❖future marking	*will* | *going to*	Travis
❖existentials	*there is* | *are*	Travis
❖deontic modality	*must* | *have to*	Travis
❖past tense	*past* | *bin* + *V*	Mailhammer
❖try	*try and* | *or V*	Matsumoto
❖ t/d deletion		Walker
❖in(g)		Travis

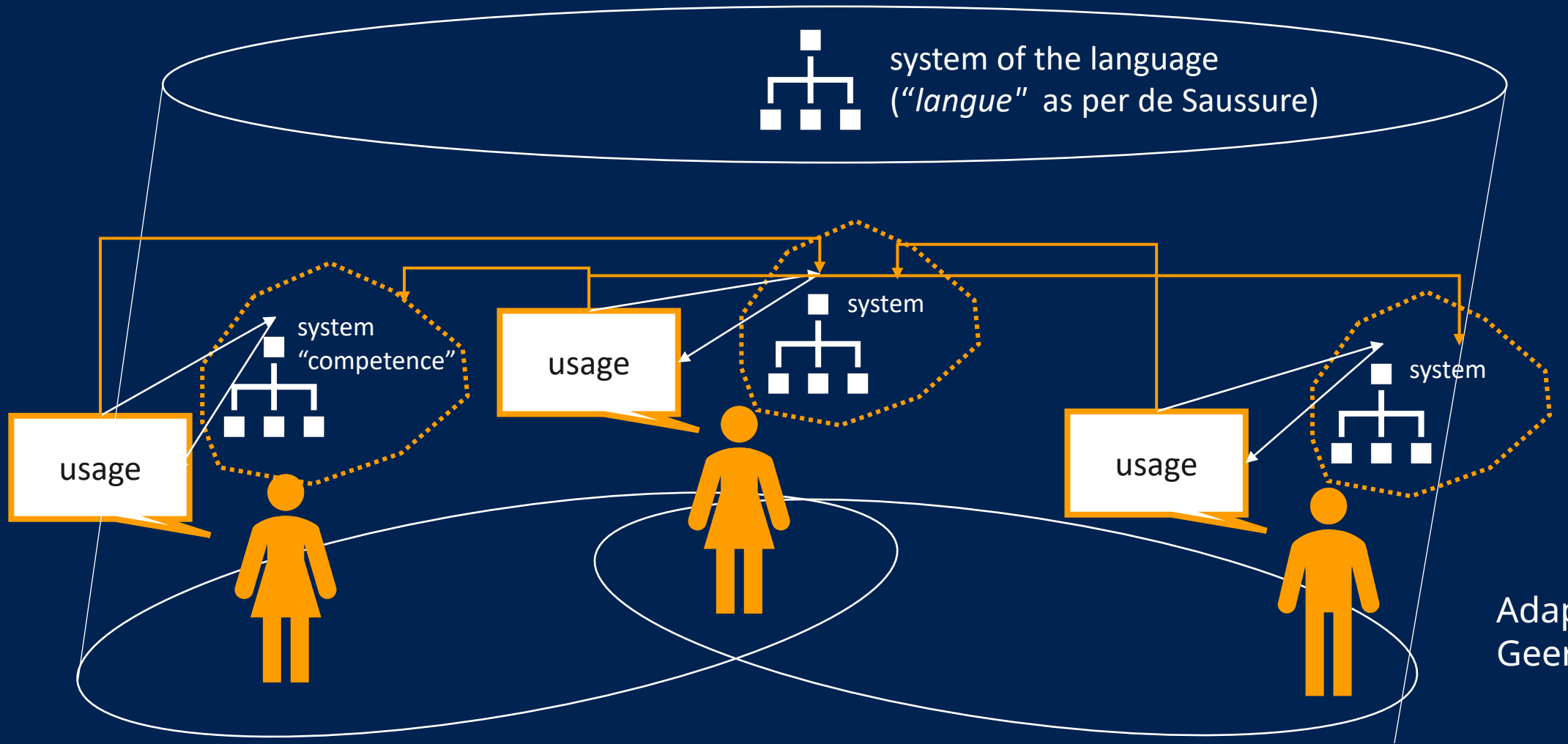
Alternate ways of saying the same thing

Variation between "alternate ways of saying 'the same' thing"
(Labov 1972: 188)

Probabilistic variation between variants that are in principle available to all members of a speech community

Intra-speaker variation

Intra-speaker variation acquired by social realignment based on actual usage



Adapted from Geeraerts 2018

The dative alternation in English

“I've never even bought a gun myself. My dad’s **given it to me** or someone's **given me one**. So I'm probably real illegal, you know, carrying guns that aren't even mine.

(Switchboard US F/SM/67)

An example from the Switchboard corpus

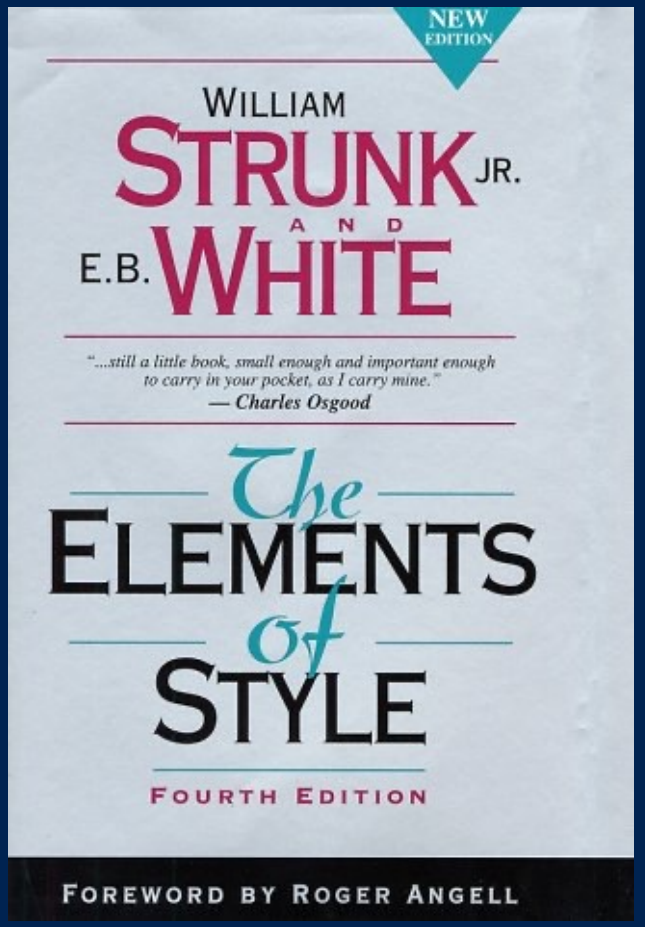
In 1 utterance both variants are used

Is a variable grammar a good thing?

If you ask prescriptive style guides, the answer is no.

> The use of *which* for *that* is common in written and spoken language.
> ("Let us now go even unto Bethlehem, and see this thing which is come to pass."). […]
> But it would be a convenience to all if these two pronouns were used with precision. Careful writers, watchful for small conveniences, go *which*-hunting, remove the defining *whiches*, and by so doing, improve their work.

(Strunk & White 1999[4th ed]: 59)



Relativizer variation in English

Subject position

(1) a. The car *that* caught fire

b. The car *which* caught fire

Object position

(2) a. The car *that* I saw

b. The car *which* I saw

c. The car ___ I saw

Is there even variation?
The principle of isomorphism

In many flavours of construction grammar the principle of isomorphism is accepted a priori. But as we just saw with *that, which,* ___ there are three possible strategies for relativizing in English.

> “The commonly accepted axiom that no true synonyms exist, i.e., that different forms must have different meanings. (Haiman 1980:516)

> “If two constructions are syntactically distinct, they must be semantically or pragmatically distinct. (Goldberg 1995: 67)

Variation persists across time

> “The relation between functionally similar forms is often described in terms of competition. […] Over time only one form can survive […] or each form must find its unique niche in functional space. […]
> If isomorphism were the only force shaping communicative codes, synonymy would be the exception in synchrony and should be consistently rooted out in diachrony. (De Smet et al. 2018: 197-198)

In other words: variable forms persist across time, but are perhaps suboptimal because they are complex.

Why grammatical variation might be difficult

Typically conditioned probabilistically by any number of contextual constraints, which need to be checked before producing a variant.

animacy, length of certain constituents, matrix verbs, phonological conditions vs. idiomaticity and chunking

The contextual scanning must happen at some level: automatic or under overt executive control. It is plausible that this extra cognitive work increases production difficulty (making it more complex).

Language complexity

Growing body of evidence showing that human languages differ in terms of complexity (Miestamo 2008)

- absolute complexity = length of grammar
- relative complexity = cost and/or difficulty of grammar for language users

→ Does absolute complexity correlate with relative complexity, as our linguistic intuition would have it?

Interim summary

Widespread (implicit) assumption that grammatical variation is messy, abnormal, short-lived diachronically, and thus suboptimal/difficult

Beyond such prejudices, it is not implausible that checking the conditioning of grammatical variants indeed increases cognitive load.

But the link between grammatical variation and increased relative complexity has never been demonstrated empirically.

At least not until…

Gardner, Matt Hunt, Eva Uffing, Nicholas Van Vaeck & Benedikt Szmrecsanyi (2021).
Variation isn't that hard: Morphosyntactic choice does not predict production difficulty.
PLOS ONE 16(6). e0252602.
DOI: 10.1371/journal.pone.0252602

Data and Methodogy

Switchboard corpus

(Godfrey et al. 1992)

- Telephone conversations (n = 2239) between 542 American English speakers (strangers) recorded by Texas Instruments in 1989/1990
- Conversations typically last for 5 minutes
- Time-aligned transcripts available for all conversations
- Disfluency and discourse annotation for a subset (1280 in the 1990s + some 400 by us), which have been analyzed to a degree
  (e.g. Shriberg 1996; Shriberg & Stolcke 1996; Wieling et al. 2016)

Demographics of the Switchboard corpus

The cut-off date is 1960

Dialect

South Midland* (155)

Western (85)

North Midland (77)

Northern (75)

Southern (56)

New York City (33)

Mixed (26)

New England (21)

Age

20-29* (140)

30-39 (179)

44-49 (112)

50-59 (87)

60-69 (13)

Sex

Female* (239)

Male (229)

Education

< High school (14)

High school < college (39)

College (309)

> College (176)

Unknown (4)

*subset reported in Gardner et al. (2021)

When looking at the entire Switchboard corpus the results are the same.

Methodology in short

Basic idea: Determine if grammatical variation contexts attract disfluencies as markers of production difficulty.

1. Tap into the Switchboard corpus and take transcripts (turns) as data points

Methodology in short

Basic idea: Determine if grammatical variation contexts attract disfluencies as markers of production difficulty.

1. Tap into the Switchboard corpus and take transcripts (turns) as data points
2. Identify and quantify filled and unfilled pauses in each conversation

Measuring processing difficulty

Operationalize production difficulty as being proportional to the extent to which production triggers disfluencies:

- Overt disfluencies: filled pauses or delay markers (*um*, *uh*)
- Silence: unfilled pauses (<50db, >130ms) in the speech stream between words or before utterances

Interpretation: utterances with **more** filled or unfilled pauses are considered to have been **harder** to produce.

Filled and unfilled pauses

Well, um, um, um, I think that uh once we get the house refinanced, we're gonna probably try to take our free tickets and either go to Cancún or do the little uh trip to Ca- Southern California and then on up to (592ms) Utah.

(female/South Midlands/born 1961)

{D well } {F um } {F um } {F um } I think that {F uh } once we get the house refinanced / we're gonna probably try to take our free tickets / {C and } either go to Cancún / {C or } do the little {F uh } trip to [Ca, + Southern California] and then on up to Utah.

(Annotated for disfluencies)

Methodology in short

Basic idea: Determine if grammatical variation contexts attract disfluencies as markers of production difficulty.

1. Tap into the Switchboard corpus and take transcripts (turns) as data points
2. Identify and quantify filled and unfilled pauses in each conversation
3. Identify and quantify variable contexts (i.e. sites where speakers have the choice between different grammatical ways of saying the same thing à la Labov (1972: 188)

Grammatical variation phenomena considered

Compiled a list of as many major alternations in mainstream North American English as possible

"Usual suspects" representing a broad spectrum of types of variation

Retrieval: hand-coding

1 **Particle placement**
2 **Dative alternation**
3 **Genitive alternation**
4 **Analytic vs synthetic comparatives**
5 ***That* vs. zero complement clauses**
6 **Infinitival vs. gerundial complementation**
7 ***That* vs. gerundial complementation**
8 **Expressions of future temporal reference**
9 **Expressions of deontic modality**
10 **Expressions of stative possession**
11 ***That* vs. wh- vs. zero relativizers**
12 ***Not* vs. *no* negation**
13 ***Not* vs. auxiliary contraction**
14 **Indefinite pronouns with singular human reference**
15 **Coordinated pronouns**
16 **Quotatives**
17 ***try and* vs *try to* vs *try -ing***
18 ***tried to* vs *tried -ing***
19 ***without any* vs. *with no-***
20 **Unrestricted Relative Pronouns**
21 ***there is* vs. *there are* for plurals**
22 ***think/hear/know/talk/speak about* vs. *of***

Methodology in short

Basic idea: Determine if grammatical variation contexts attract disfluencies as markers of production difficulty.

1. Tap into the Switchboard corpus and take transcripts (turns) as data points
2. Identify and quantify filled and unfilled pauses in each conversation
3. Identify and quantify variable contexts (i.e. sites where speakers have the choice between different grammatical ways of saying the same thing à la Labov (1972: 188)
4. Check if there is a correlation.

Techniques

- Exploratory analyses with repeated measures correlation

- Random forests and mixed-effects regressions
  - (un)filled pauses ~ all variables + three control variables
    → mean word length
    speech rate (words / s)
    turn duration
  - (un)filled pauses ~ absence/presence of all variables + control vars

    (simplest model, Occam's razor)

Results

Some numbers

In over 240 hours of recording,

- …there are about 34 hours of silence (**unfilled pauses**)
- …there are 91,001 *um*'s (n = 21,187) and *uh*'s (n = 69,814) (**filled pauses**)
- …there are 62,329 **grammatical variable contexts**

Exemplification

Well, um, um, um, I think that uh once we get the house refinanced, we're gonna probably try to take our free tickets and either go to Cancún or do the little uh trip to Ca- Southern California and then on up to (592ms) Utah.

(female/South Midlands/born 1961)

5 filled pauses

592 ms of unfilled pauses

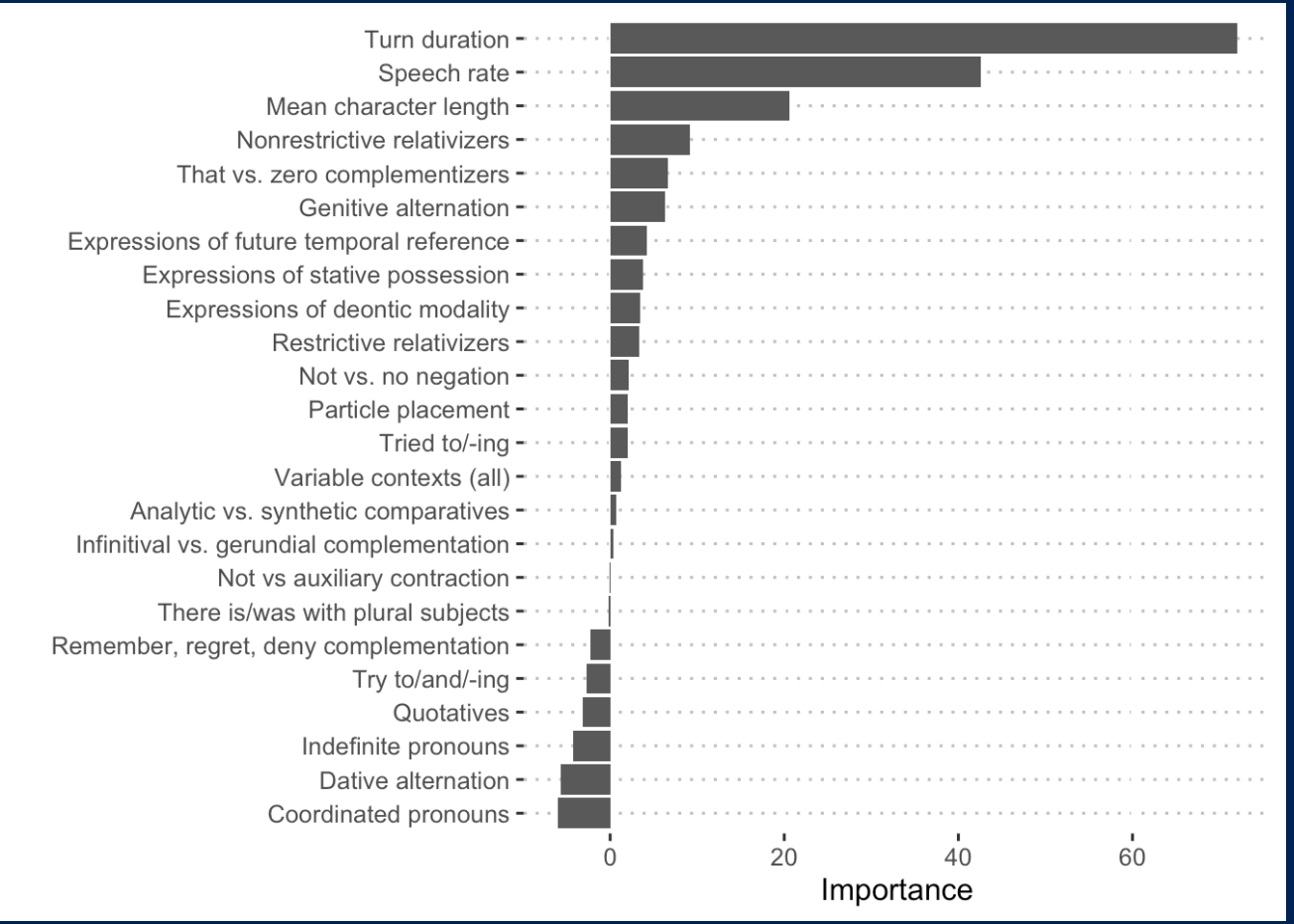
→ dependent variables

Filled pauses
(delay markers)

Regression take-away

Longer turns → more disfluencies
Higher speech rates → fewer disfluencies
Longer words → fewer disfluencies

The simplest models (Occam's razor) have absence/presence as levels for the variables (future temporal reference***, not/no*, deontic modality*)

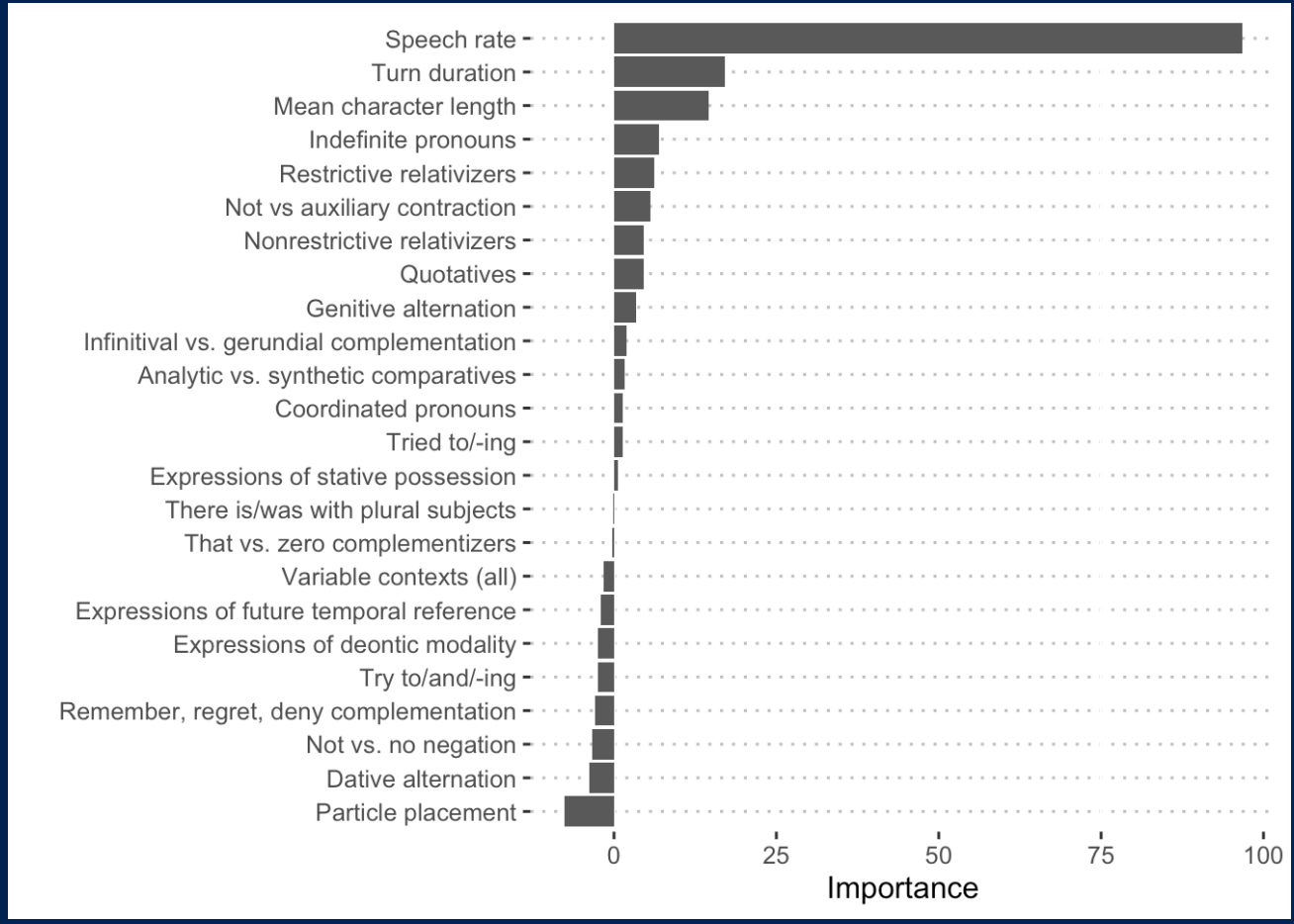


Unfilled pauses
(silence)

Regression take-away

Longer turns → fewer disfluencies (?)
Higher speech rates → fewer disfluencies
Longer words → fewer disfluencies

The simplest models (Occam's razor) have absence/presence as levels for the variables (quotative***, remember/regret/deny***, nonrestrictive relativizer**,not/no*, indefinite pronouns*, coordinated pronouns*)



In other words…

There is no evidence whatsoever that higher frequencies of variable contexts (not even with a binary absence/presence) correlate positively with disfluencies, as they should if variation indeed impedes production

So, **grammatical variation does not appear to be particularly difficult!**

Theoretical implications

1. **Isomorphism and No Synonymy**: we knew before that variation is pervasive and not necessarily short-lived. Now we also know that it doesn't cause production problems for speakers. It is time to re-think some dogmas.
2. **Diachronically stable grammatical alternations**: these are not so mysterious after all – alternations can be stable because variation is not suboptimal.
3. **Complexity theory**: here we have one instructive case where increased absolute complexity (length of grammar) does not correlate with increased relative complexity (cost of production).

Current and future work

(i.e., since Thomas Van Hoey (that's me) joined in January this year)

❖ Beyond canonical disfluencies

We are currently annotating for more disfluencies:

- Fillers — *uh, um, huh, oh*
- Editing terms — *I mean, sorry, excuse me*
- Discourse markers — *you know, well, actually, so, like*
- Coordinating conjunctions — *and, and then, but, because*
- Asides
- Restart and repair — [Ca-, Southern California]

Current and future work

(i.e., since Thomas Van Hoey (that's me) joined in January this year)

- Probabilistic modelling

Some choices are hard to make than others. So when the model predicts a chance of 0.5 this is a harder choice (between two variants) than when it is 0 or 1.

We are working on integrating this into this line of research.

Current and future work

(i.e., since Thomas Van Hoey (that's me) joined in January this year)

- Stable versus dynamic variables/alternations
  Or the longevity of certain alternations

- Number of probabilistic constraints involved

If it matters, then it will probably be that variables with many constraints (i.e., more cognitive work to be taken into account) have more / longer disfluencies.

- Also in other languages?

In sum

There is a robust finding (Gardner et al. 2021) that true variable contexts are not necessarily suboptimal in terms of relative complexity, not for the subset of the Switchboard corpus nor for the whole Switchboard corpus.

But at the same time there are still many aspects of the data that were left unexplored. We are working on making our way through the data so that we can map these angles and further investigate whether the isomorphists were right.

Thanks and happy winter solstice, mate!

Thomas Van Hoey

Matt Hunt Gardner

Benedikt Szmrecsanyi

Funding
FWO G.0C59.13N
KU Leuven C1 3H220293